Filter Detection using a VGG16 based neural network

Tarshen Naidoo

November 2021

Abstract

Technology has improved substantially around image processing in recent years and the widespread adoption on social media networks has led to trust issues among social media user-bases - especially young users - with regards to personal reflection[8] and towards others. Filters are generally used to remove flaws in one's image and increase potential attraction or value in a viewer's mind. With filters being increasingly abundant on social media platforms - especially with females [6], the mental health impact on users becomes increasingly severe. [5]

This report will focus on the ability to detect filter presence and the type of filter used based on a set of pre-defined filters. A neural network based on the VGG16 CNN model [11] will be used alongside a database of face images with Adobe Photoshop Express image augmentations. Success metrics include confusion matrices.

Introduction

Since the integration of visual mediums on social media platforms, filters have been a popular tool to increase a user's potential appeal. Filters are generally used to remove flaws and increase appeal. Filters have a more general uses such as comedy or experimental filters out of interest, however, filters with the primary function of beauty some of the most used.

In this paper, we will look into filters that do not add arbitrary images and visuals but apply morphological operations on source images. Filters that will be used are from the Adobe Photoshop Express application. Filters used are 'clarity', 'dehaze', 'grain', and 'vignette' filters.

A VGG16 based convolutional model will be used to detect the presence of these filters.

Following positive results, possible follow up research could involve filter reversal on certain filters. This has various benefits in the realm of social media and other areas where filter remove would be beneficial.

The dataset to be used will be Japanese Female Facial Expression dataset containing images of a combination of different Japanese females, facial expressions and variants. [10] [9] There are 212 images in this dataset.

This dataset is further augmented with four different filters. The total images to be used for this research is 1060 images.

Background

Filters

Instagram were involved in the early pioneering of social media filters as a way to improve the visual fidelity of images taken through smartphone lenses. As these lenses were still being improved and were nowhere near the quality they are now; these filters were a means to add visual appeal by dramatically altering the way images looked. These early filters were not very subtle and gave images a certain 'sameness' as they were almost entirely fixed. [7]

At some point around 2014, these types of filters fell out of favour with the general user-base as smartphone camera qualities improved to satisfactory levels. It was at this point that the new types of filters - designed to subtly mask an image - gained popularity for their ability to create an almost unrealistic, yet believable, image. As these types of filters were refined further, the ability for people to identify the presence of filters became increasingly difficult.

VGG16

The VGG16 Convolutional Neural Network is a CNN developed by the Visual Geometry Group from Oxford University. The CNN was used in an image net competition in 2014 where it performed well. Since then, it has been refined and used in numerous different applications including malicious software classification [7] and Kiwi fruit detection [3]

The current model is a fusion of a 16-layer and 19-layer CNN. This fusion model achieves 7.1% and 7% on the ILSVRC-2012 top-5 classification error[11] with success rates between 86% and 93% on different datasets.

1 Research Hypothesis

"To what degree of accuracy can the proposed Neural Network model achieve on the testing dataset in identifying the correct filter?".

To answer this question; the architectural design of the CNN must be considered.

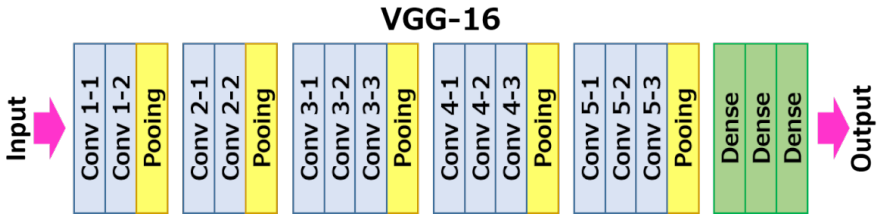

Figure 1: VGG16 CNN architecture

The CNN is designed to incrementally build to a higher level of abstraction to identify more general trends. This is accomplished through the max pooling layer. The output of this VGG16 model will feed into a final classification layer. Where the shape is a vector of size 5 - indicating no filter or a specific filter (one of four).

Deep CNN models tend to generalise well to new datasets after a long enough period of training. Thus, given training time, it is feasible for the proposed model to achieve a high accuracy

Research Methodology

The original JAFFE dataset has been used. The 212 images have had numerous different filters applied to them. The algorithm will read in these images and labels and split them into training, validation, and testing sets of splits 0.7, 0.15, and 0.15 respectively.

The model will be trained and validated on the respective data until convergence.

As this is a classification problem. A confusion matrix will be the best way of assessing the accuracy of the data.

Results

The model took 16 iterations totalling 89.4 minutes to run. The algorithm slowly trained to 60% accuracy before quickly jumping to 95% accuracy on the validation data. The confusion matrix below shows that the model performed very well under the test data provided achieving 98.7% accuracy which correlates well with the initial hypothesis.

Predict/Actual	Original	Clarity	Dehaze	Grain	Vignette
Original	25	0	0	0	1
Clarity	0	31	0	0	0
Dehaze	0	0	40	0	0
Grain	0	0	1	21	0
Vignette	0	0	0	0	40

Confusion matrix for filter prediction

While this seems a very positive result; the dataset is only 1060 images large and the algorithm should be implemented to include more variation within images and filters. As a result the model could very well be over-fitting to the data provided. Convolutional neural networks are known for overfitting. [1]

Research put into detecting Instagram filtered images is relatively new as the prominence of subtle Instagram filters are a relatively new emergence. That said, there has been research done in this space with regards to feature destylization. [4] Instead of modelling to recognize each filter individually, the model proposed analyzes each image for features that more closely resemble an image's original colour.

These results demonstrates a foundation from which to explore the addition of more filters, and more variation in source images. The VGG16 CNN demonstrate the ability for CNNs to generalise well and the implementation of the VGG16 model and/or the adaptation to new datasets with fine tuning could lead to more general results.

1.1 Conclusion

There has been a growing level of mental health concerns amongst those that use social media, particularly during Covid 19. [2] While filters may not be objectively malicious; it is evident that there is a need to develop the technology to identify and possibly reverse them.

The results from this report demonstrate the feasiblity for the VGG16 model as a foundation for filter detection. It is possible that future research may look into reconstructing the unfiltered image from the detected filter and the dropout layers of the VGG16 model.

References

[1] Andrei Dmitri Gavrilov et al. *Preventing Model Overfitting and Underfitting in Convolutional Neural Networks*. URL: https://www.researchgate.net/profile/Jack-Deng-2/publication/331218117_Preventing_Model_Overfitting_and_Underfitting_in_Convolutional_Neural_Networks/links/5f9eb3ac299bf1b53e5653e7/Preventing-Model-Overfitting-

and-Underfitting-in-Convolutional-Neural-Networks.pdf. (accessed: 12/11/21).

[2] Junling Gao et al. *Mental health problems and social media exposure during COVID-19 outbreak*. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0231924. (accessed: 12/11/21).

[3] Liu et al. *Improved Kiwifruit Detection Using Pre-Trained VGG16 With RGB and NIR Information Fusion*. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8943411. (accessed: 12/11/21).

[4] Zhe Wu et al. *Recognizing Instagram Filtered Images with Feature De-stylization*. URL: https://arxiv.org/pdf/1912.13000.pdf. (accessed: 12/11/21).

[5] Mike Conway and Daniel O'Connor. *Social Media, Big Data, and Mental Health: Current Advances and Ethical Implications*. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4815031/. (accessed: 11/11/21).

[6] Sarah Fielding. *90% of Women Report Using a Filter on Their Photos*. URL: https://www.verywellmind.com/90-of-women-report-using-a-filter-on-their-photos-516048. (accessed: 11/11/21).

[7] Kate Imbach. *A Visual History of the Instagram Filter*. URL: https://medium.com/@kate8/a-visual-history-of-the-instagram-filter-f1ffe8168091. (accessed: 11/11/21).

[8] Christine Lavrence and Carolina Cambre. *'Do I Look Like My Selfie?'* URL: https://journals.sagepub.com/doi/pdf/10.1177/2056305120955182. (accessed: 11/11/21).

[9] Lyons. *'Excavating AI' Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset*. URL: https://arxiv.org/pdf/2107.13998.pdf. (accessed: 12/11/21).

[10] Gyoba Lyons Kamachi. *Coding Facial Expressions with Gabor Wavelets (IVC Special Issue)*. URL: https://zenodo.org/record/4029680#.YY2hcp5BxD8. (accessed: 12/11/21).

[11] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Visual Recognition*. URL: https://www.robots.ox.ac.uk/~vgg/research/very_deep/. (accessed: 12/11/21).